

METHOD AND APPARATUS FOR LEARNING A PROBABILISTIC GENERATIVE MODEL FOR TEXT

ABSTRACT

One embodiment of the present invention provides a system that learns a generative model for textual documents. During operation, the system receives a current model, which contains terminal nodes representing random variables for words and cluster nodes representing clusters of conceptually related words. Within the current model, nodes are coupled together by weighted links, so that if a cluster node in the probabilistic model fires, a weighted link from the cluster node to another node causes the other node to fire with a probability proportionate to the link weight. The system also receives a set of training documents, wherein each training document contains a set of words. Next, the system applies the set of training documents to the current model to produce a new model.